I loved the way all the four practical exercises prepared and cleaned the data for mining. I was already aware of most of the techniques like:

1. Dropping null values or replacing the null values with mean/ median/ min/ max (whichever is applicable)
2. Making sure the datatypes are accurate for each column
3. Converting datetime to date, year, month, day, hour, minutes, seconds (whichever combination is applicable) if necessary
4. Convert all columns with object datatypes to either integer or float datatype for training dataset
5. Dropping unnecessary columns and/ or columns with mostly null values from the dataset
6. Using correlation between predicted variable and explanatory variable(s) to identify important features
7. Converting qualitative variables to categorical feature, also convert string categories to numerical categories for training
8. Making sure that key columns are properly formatted while merging or joining two or more datasets
9. Updating column names if necessary
10. Manipulation of string using string functions or regex to weed out bad or special characters
11. Filtering out records with bad values
12. Standardizing universal values, such as state names, ZIP codes, country names etc.
13. Dropping outliers using boxplot on each important numerical feature if necessary.

What I have learned is that I could also incorporate graphs/ visualization as part of exploratory data analysis. Going forward I will be using the following techniques:

1. Using heatmap for correlation between all the explanatory variables
2. Using pair plots to understand the relationship characteristics between response variable and explanatory variables
3. Using histograms on each feature to understand the distribution

I did not face any big challenges understanding the steps related to data preparation and data cleansing process. But I do face difficulties when it comes to writing lambda functions and understanding the difference between supervised learning and unsupervised learning for k means clustering.

The most rewarding thing to learn from these exercises for me was to learn the significance of correlation heatmap in the feature selection process.